S&DS 230 Final Project: Gun Ownership, Violent Crime, and Gun Control Policy in the United States

Due: 2019 December 08

Author: Megan Zhang

Discussants:

Articles mentioned, with link included:

CBS News on Mass Shootings This Year

CNS News on More Guns, Less Violence?

A Blog Article on the Link between Gun Ownership and Deaths

The Pew Research Center on Gun Deaths in the U.S.

Harvard Injury Control Research Center on Availibility of Guns and Homicides

Introduction

In 2019 to date, according to the Gun Violence Archive, there have been more mass shootings than days in the year. (a mass shooting is defined as any incident in which at least four people are shot. (Link to news article.)). The debate on gun control policy in the United State has become increasingly relevant and heated, with sharp divides in opinion, and usually along party lines. It seems intuitive to some that controlling gun access will decrease rates of violent crime across the country. On the other hand, some claim that controlling gun access only leads to use of other weaponry for crime and that they are needed for self-defense. Interestingly, CNSNews shows (in this article) that from 1993 to 2013, gun homicide rates in the United States actually decreased, while the number of guns per person increased. However, sites such as the Pew Research Center choose to focus on the past decade as a period of rapid increase in gun deaths (link) rather than focusing on the overall decrease since the 1990s. With all of these opinions in mind, I decided to analyze the relationship between gun ownership and various crime metrics, including homicides and violent crime rates. Secondly, I will be looking at the specific gun laws of each state and comparing these metrics in order to determine whether stricter laws really do have an impact on crime rates.

The data I am using is compiled from several different sources onto the world population review website. Statistics for Gun Deaths by State (link) are originally from the CDC's records of Firearm Mortality by State. Rates are alll recorded in deaths per 100k people. Data for Gun Laws by State (link) is compiled by GunsToCarry. Gun Ownership by State (link) is based on 2017 gun registration statistics from the Federal Bureau of Alcohol, Tobacco, Firearms, and Explosives. Homicide and Violent Crime Rate by State (link) is reported from the FBI's Uniform Crime Reports from 2017.

It does seem that many analyses on gun violence data have already been performed, to varying results. One writer claims that there is minimal correlation between gun ownership and gun violence deaths (link). These analyses by the Harvard Injury Control Research Center (link) seem to support the notion that the increased availability of guns was correlated with increased homicides. Although I cannot reach the same level of depth of analysis as these studies, my goal in this analysis is to identify possible correlations or lack thereof.

Results

Data wrangling: Combinining cleaned data into one data frame

```
lawdat <- read.csv("gun_laws_by_state.csv")
ownershipdat <- read.csv("gun_ownership_by_state.csv") %>% select(State, gunOwnership)
homicidedat <- read.csv("homicide_rate_by_state.csv") %>%
  select(State, homicideRate2017, firearmDeathRate)
violentcrime <- read.csv("violent_crime_by_state.csv")

#Joining the data using merge() and left_join
completedat <- merge(ownershipdat, homicidedat, by ="State") %>%
  left_join(violentcrime, by = "State") %>%
  merge(lawdat, by = "State")
completedat <- completedat %>% rename(homicideRate = homicideRate2017,
                                      violentcrimeRate = violentcrime_rate)
#Converting to narrow form for visualization
completedat_narrow <- tidyr::pivot_longer(completedat,
                  cols = c(firearmDeathRate, homicideRate, violentcrimeRate))
completedat_narrow <- completedat_narrow %>% rename(crimeType = name,
                                                    rate = value)
```

Not much data cleaning was required because the calculated statistics (crime rate per capita, percentage of adults with gun ownership) were already made available by the World Population Review and its sources. I downloaded four separate datasets relating to gun ownership, firearm deaths, crime rates by state, and gun laws by state, and edited them in Excel as .csv files to make sure the variable names were appropriate. Then, I joined them all to one dataset using the State variable as the key. I also created a narrow version of the dataset to be used for visualization.

Visualizing the data: Comparing crime metrics against Gun Ownership and Gun Laws using ggplot

All code is listed in the Appendix.

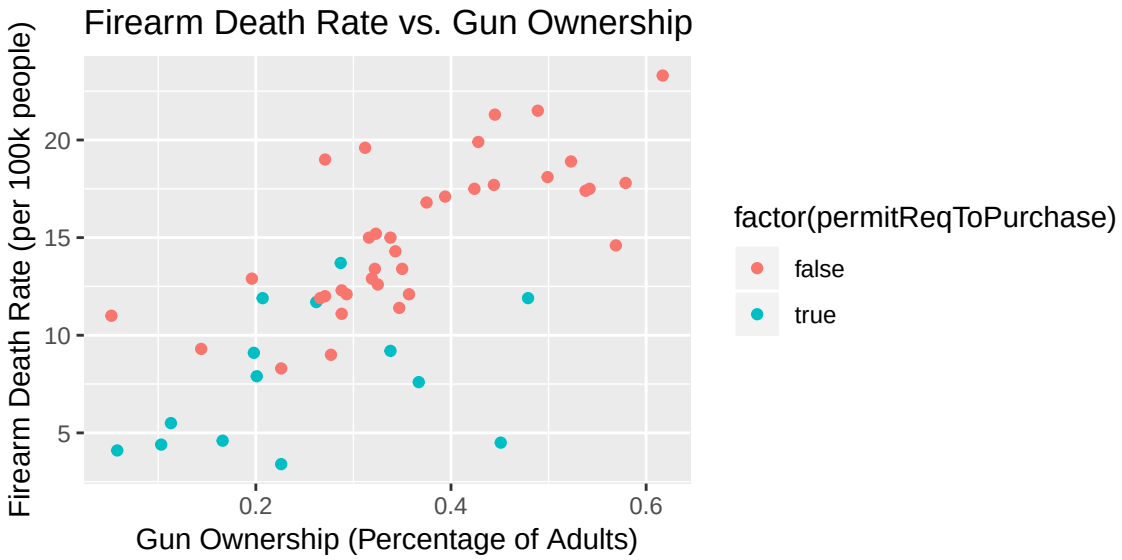


Figure 1. The above plot depicts the relationship between gun ownership (measured in percentage of adults who own a gun) and the Firearm Death Rate for all fifty states in the U.S. As we can see there does seem to be a positive correlation, which we will analyze further in the next section. I have also separated the points according to factor which is whether or not the state requires gun purchasers to have a permit before they may purchase a gun. I chose this as the factor because of all the variables relating to gun law which we were given, a permit to purchase is considered one of the first steps in the process of purchasing a gun, and would present the first barrier to access. Additionally, the other factors relating to law were less commonly present throughout the states.

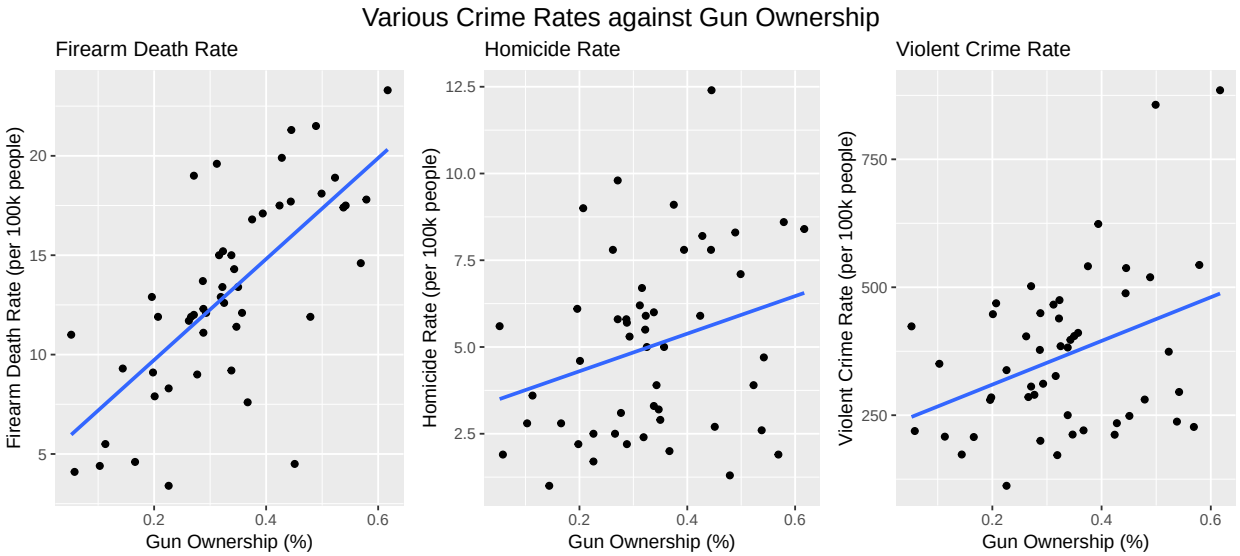


Figure 2. This figure shows side-by-side plots of different crime metrics against gun ownership: First, firearm death rate (which is not officially a measure of crime), homicide rate, and violent crime rate. I chose to depict these three separately instead of using a factor variable because the scales were quite different (notice violent crime rates are much higher than the other two). Fitting a linear regression to each plot seems to show positive correlation for all three metrics; however, looking at the plotted points there really does not seem to be a clear correlation between gun ownership and homicide or violent crime rates.

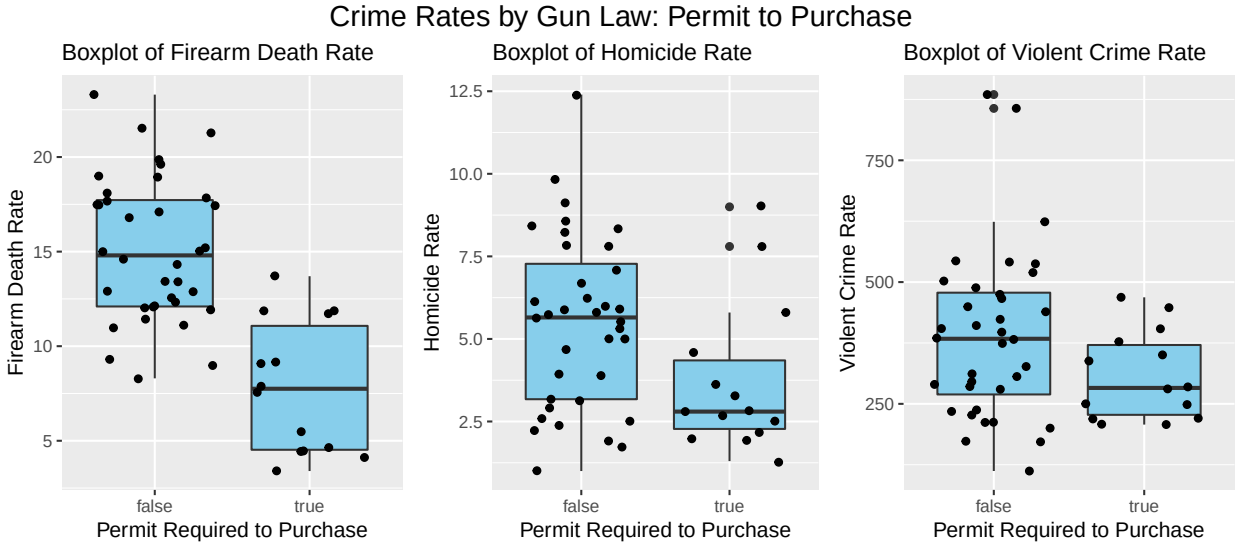


Figure 3. This figure shows side-by-side boxplots of each crime/death metric, divided according to whether

or not the state requires a permit to purchase firearms.

Discussion Unsurprisingly, we can see that states with more relaxed gun control laws have higher rates of gun ownership, shown by the factor separation in Figure 1; additionally, these seem to be correlated with higher gun death rates. Figure 2 seems to show a positive correlation between gun ownership and each of the metrics used; however, it does appear that using linear regression may not be the best choice for analysis since the points are relatively scattered. From Figure 3, it does appear that states which require a permit to purchase a firearm have lower Firearm Death Rate, Homicide Rates, and Violent Crime Rates. These differences will be explored in the Analysis section.

Analysis: Linear Models, Hypothesis tests of means, Confidence Intervals for Correlation and Regression Slopes

1. Fitting Linear Models for the Three Metrics

Model for firearm Death Rate vs Gun Ownership. The slope coefficient is extracted from each linear model.

```
# Models a Linear Regression for firearm death rate vs. gun
# ownership
lm_1 <- lm(firearmDeathRate ~ gunOwnership, data = completedat)
# Extracts the slope coefficient
summary(lm_1)$coefficients[2, ]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 2.539908e+01 3.762385e+00 6.750791e+00 1.772994e-08
```

Homicide Rate vs Gun Ownership

```
lm_2 <- lm(homicideRate ~ gunOwnership, data = completedat)
summary(lm_2)$coefficients[2, ]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
## 5.41340074 2.68992693 2.01247130 0.04979931
```

Violent Crime vs. Gun Ownership

```
lm_3 <- lm(violentcrimeRate ~ gunOwnership, data = completedat)
summary(lm_3)$coefficients[2, ]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 4.267615e+02 1.567170e+02 2.723135e+00 8.989169e-03
```

Discussion. Here I fitted linear regression to each of the three plots from Figure 2. We can see the resulting linear regression in blue on each graph. All three gave slope values greater than 0. If we take the following null and alternative hypothesis:

$$H_0 : \beta_1 = 0; H_A : \beta_1 \neq 0; \alpha = 0.05$$

We find that all three models give positive slopes with p-value less than α, signifying a positive relationship. For firearm deaths, for example, each additional percentage in gun ownership appears to increase firearm

death rate by about 25 per 100k people. This increase is 5.14 per 100k for homicides, and 427 for violent crimes.

However, only the first model gives us a remotely high **r-squared value** of **0.487**, while the other two are very low (**0.078** and **0.134**). These values were calculated and extracted from the lm() function. Fitting **polynomials** up to degree 5 does give us higher r-squared and adjusted r-squared values (not shown in the code chunk above); however, I believe using this method will result in overfitting and looking at the data plotted above, it does not seem reasonable that any other degree polynomial would be a good fit for the data. Next, we'll create confidence intervals for correlation using the bootstrap method to see if the variables are actually correlated.

2. Confidence Interval for Correlation

We examine ρ, the correlation constant, where $\rho = 0$ signifies no correlation. We will create 95% confidence intervals for correlation between each of the three variables and gun ownership using the bootstrap method.

```
#Creating correlation interval using the Bootstrap Distribution
bootstrap_dist <- NULL
for (i in 1:10000){
  one_bootstrap_data_frame <- completedat[sample(1:50, 50, replace = TRUE), ]
  bootstrap_dist[i] <- cor(one_bootstrap_data_frame$firearmDeathRate,
  one_bootstrap_data_frame$gunOwnership)
}

corint <- quantile(bootstrap_dist, c(0.025, 0.975))
```

Confidence interval for ρ for firearm Death Rates and gun ownership : [0.5023991, 0.8465094]

Using the same code as above for the other two metrics (see appendix for full code) we arrive at:

Confidence interval for $\rho_{Homicide}$: [-0.0073454, 0.5335258].

Confidence interval for $\rho_{violentcrime}$: [0.0436859, 0.6034313].

As we can see, the confidence interval for correlation between homicide and gun ownership contains 0, suggesting these two variables may not actually be correlated. The interval for Violent crime and gun ownership does not contain zero and implies a positive correlation; however, the interval is very close to 0, which I believe warrants further analysis on the correlation between these two variables. The confidence interval for ρ of firearm death rates and gun ownership was high and did not contain 0.

3. Confidence Interval for Regression Slopes

This confidence interval is using the bootstrap method. Since the sample size is small, and the data does not seem to come from a normal distribution (the variation in states is quite large), the bootstrap seems to be a more accurate method of estimation.

Bootstrap Confidence interval for Slope of Firearm death rate vs. Gun Ownership:

```
#Using the bootstrap to create a confidence interval for slope
nrep <- 10000
n_cases <- dim(completedat)[1]
result_vec <- rep(0, nrep)
for (i in 1:nrep){
  boot_sample <- dplyr::sample_n(completedat, size = n_cases, replace = TRUE)
  boot_fit <- lm(firearmDeathRate ~ gunOwnership, data = boot_sample)
  result_vec[i] <- coef(boot_fit)[2]
}

slopeconfint <- quantile(result_vec, c(0.025, 0.975))
```

The confidence interval is `slopeconfint` = [17.8786945, 32.6178613].

Using the same code as above (listed in the appendix) I calculated confidence intervals for the slopes of the remaining two metrics.

Bootstrap Confidence interval for Slope (β_1) of Homicide Rate vs. Gun Ownership: The interval is [-0.1224777, 10.849594].

Bootstrap Confidence interval for Slope (β_1) of Violent Crime Rate vs. Gun Ownership (code omitted): The interval is [42.2181056, 830.0721187].

Notice that the confidence interval for the slope of Homicide Rate vs. Gun Ownership contains 0, which implies there is not a significant linear relationship between the two variables.

4. Comparison of Means: T test and Permutation Test

In this section we compare the means on our crime-related metrics based on different laws established by the state. The first is whether or not the state requires a permit to purchase firearms. We'll run a t-test first in order to compare results of different tests.

```
#Separating the dataset on whether a permit is required to purchase
permit_req <- completedat %>% filter(permitReqToPurchase == "true")
permit_notreq <- completedat %>% filter(permitReqToPurchase == "false")

t.test(permit_req$firearmDeathRate,
       permit_notreq$firearmDeathRate, alternative = "less")$p.value
```

```
## [1] 4.131293e-07
```

```
t.test(permit_req$violentcrimeRate,
       permit_notreq$violentcrimeRate, alternative = "less")$p.value
```

```
## [1] 0.01866666
```

```
t.test(permit_req$homicideRate,
       permit_notreq$homicideRate, alternative = "less")$p.value
```

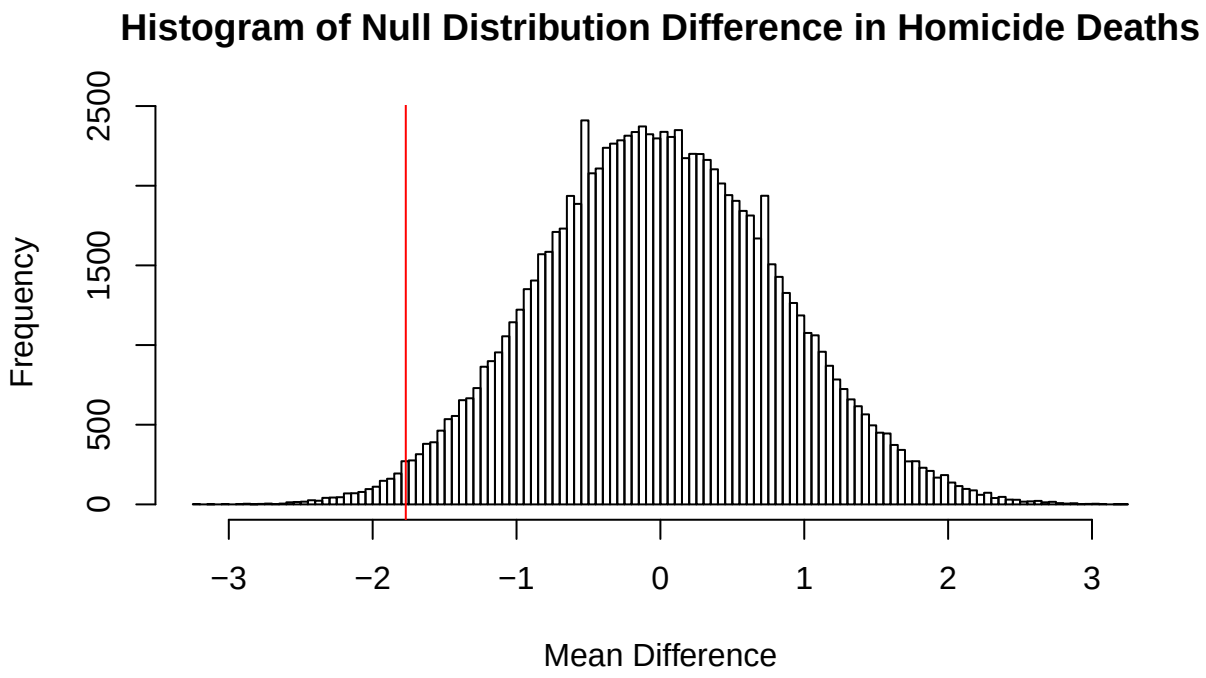
```
## [1] 0.01306914
```

I used the standard t.test() function to find p-values for comparison of each crime metric dependent on whether or not a permit in that state is required to purchase a gun. I used a one-tailed test and found the p values were all less than $\alpha = 0.05$, showing there is a significant negative difference between the means (the mean crime rate for states which require a permit is less than for states not requiring one). However, looking at distributions of the data reveals the points are not very normal; additionally, our sample size is not large, so it seems there are some concerns in performing a t-test. I think a more accurate test in this case would be a permutation test, which is what we will explore next.

The following is an example of a permutation test using the data for Homicide Death Rates.

```
obs_stat <- mean(permit_req$homicideRate) - mean(permit_notreq$homicideRate)
combined <- c(permit_req$homicideRate, permit_notreq$homicideRate)
null_dist <- c()
for (i in 1:100000){
  shuff_data <- sample(combined)
  shuff_req <- shuff_data[1:14]
```

```r
  shuff_notreq <- shuff_data[15:50]
  null_dist[i] <- mean(shuff_req) - mean(shuff_notreq)
}
hist(null_dist, main = "Histogram of Null Distribution Difference in Homicide Deaths",
  xlab = "Mean Difference",
  ylab = "Frequency",
  nclass = 100)
abline(v = obs_stat, col = "red")
```




```r
#p-value for the observed Difference in mean
(p_val <- sum(null_dist <= obs_stat)/100000)
```

```
## [1] 0.01353
```

I used the same method to calculate a p-value for the observed difference in means for the other two crime metrics (Histograms omitted, and code included in the appendix).

P-value for difference in means observed for Violent Crime Rates:

```
## [1] 0.04599
```

P-value for difference in means observed for Firearm Death Rates:

```
## [1] 0
```

Conclusion

In this analysis we were able to reach several interesting results. Firstly, there is positive correlation and linear relationship between gun ownership and firearm deaths. It seems fairly intuitive that the rate of gun deaths would increase as gun ownership also increases, so this result is not surprising to me. It seems obvious that if there are more guns available, there will be more deaths which result from guns, whether by murder or by accident. However, political discussions diverge on how useful this statistic actually is. According to the Pew Research Center, sixty percent of gun deaths in 2017 were suicides while 37% were murders. Some argue that the fact that most gun deaths are suicides means firearms are not linked to higher murder rates and thus additional regulation is not warranted. To examine this I looked at other metrics of crime and death such as homicide rates and violent crime rates, which I believe are more accurate metrics of how dangerous a city or place is.

Surprisingly, I found that there does not seem to be statistically significant correlation between homicide rates and gun ownership, nor did there seem to be a significant linear relationship between the two variables. Additionally, while there seems to be a positive linear relationship between violent crime and gun ownership, the correlation between the two variables is low. However, one result which was consistent across all three metrics was that when looking at gun laws for each state, there was a statistically significant negative difference between crime rates and deaths in states that required permits as opposed to those that did not. It seems that on average, rates of violent crime, homicides, and firearm deaths are lower in states that require a permit to purchase firearms. I believe further study on this subject will be required in order to better understand the relationship between gun ownership, crime, and depths. The dataset I used for this project was very limited, in that it only contained one year's worth of data for only fifty states, which had much variation. Perhaps examining the same statistics over the past ten years would provide more comprehensive results; or, perhaps using the same variables but examining on a smaller scale such as cities or neighborhoods would be more helpful. Unfortunately, I tried but was unsucessful at finding such statistics.

Reflection

This project probably took me about 15 hours total. However, I spent at least 5 hours searching for a data set at the beginning of this project. I initially planned to do an analysis on the factors (e.g. demographic factors such as race, education, etc.) that influence opinions of American voters on gun control. However, this proved difficult because the dataset I was using consisted of mostly categorical variables, and it was difficult to use the methods we learned this semester to perform my analysis. When doing this project, I initially wanted to also analyze other gun laws, such as states which require registration or those that require background checks. I initially included a boxplot comparison of each crime metric according to whether or not the state required firarm registration, and intended to run tests to compare the means. There were several problems with this, however: Firstly, the number of states with each specific law was low, and the sample variation was too large to acheive much result. Also, not all laws were able to be described in simple true/false terms as the variables I used were. In the future I hope to learn more statistical methods so I will be able to analyze data that is of interest to me despite their limitations.

Appendix

Data Visualization

```
#Figure 1
ggplot(completedat,
       aes(x = gunOwnership, y = firearmDeathRate, col = factor(permitReqToPurchase))) +
  geom_point() +
  xlab("Gun Ownership (Percentage of Adults)") +
  ylab("Firearm Death Rate (per 100k people)") +
  ggtitle("Firearm Death Rate vs. Gun Ownership")

#Figure 2
plot1 <- ggplot(completedat, aes(x = gunOwnership, y = firearmDeathRate)) +
  geom_point() +
  xlab("Gun Ownership (%)") +
  ylab("Firearm Death Rate (per 100k people)") +
  ggtitle("Firearm Death Rate") +
  theme(plot.title = element_text(size=12)) +
  geom_smooth(method = "lm", se = FALSE)

plot2 <- ggplot(completedat, aes(x = gunOwnership, y = homicideRate)) +
  geom_point() +
  xlab("Gun Ownership (%)") +
  ylab("Homicide Rate (per 100k people)") +
  ggtitle("Homicide Rate") +
  theme(plot.title = element_text(size=12)) +
  geom_smooth(method = "lm", se = FALSE)

plot3 <- ggplot(completedat, aes(x = gunOwnership, y = violentcrimeRate)) +
  geom_point() +
  xlab("Gun Ownership (%)") +
  ylab("Violent Crime Rate (per 100k people)") +
  ggtitle("Violent Crime Rate") +
  theme(plot.title = element_text(size=12)) +
  geom_smooth(method = "lm", se = FALSE)

grid.arrange(plot1, plot2, plot3, ncol=3,
             top = textGrob('Various Crime Rates against Gun Ownership',
                            gp = gpar(fontsize = 15)))

#Figure 3
plot4 <- ggplot(completedat, aes(x = factor(permitReqToPurchase), y = firearmDeathRate)) +
  geom_boxplot(fill = "skyblue") +
  labs(title = "Boxplot of Firearm Death Rate",
       x = "Permit Required to Purchase", y = "Firearm Death Rate") +
  theme(plot.title = element_text(size=12)) +
  geom_jitter()

plot5 <- ggplot(completedat, aes(x = factor(permitReqToPurchase), y = homicideRate)) +
  geom_boxplot(fill = "skyblue") +
  labs(title = "Boxplot of Homicide Rate",
       x = "Permit Required to Purchase", y = "Homicide Rate") +
  theme(plot.title = element_text(size=12)) +
```

```
  geom_jitter()

plot6 <- ggplot(completedat, aes(x = factor(permitReqToPurchase), y = violentcrimeRate)) +
  geom_boxplot(fill = "skyblue") +
  labs(title = "Boxplot of Violent Crime Rate",
       x = "Permit Required to Purchase", y = "Violent Crime Rate") +
  theme(plot.title = element_text(size=12)) +
  geom_jitter()

grid.arrange(plot4, plot5, plot6, ncol=3,
             top = textGrob('Crime Rates by Gun Law: Permit to Purchase',
                            gp = gpar(fontsize = 15)))
```

Analysis

```
#Omitted bootstrap distribution code for Correlation Interval
#Creating correlation interval using the Bootstrap Distribution
set.seed(1)
bootstrap_dist <- NULL
for (i in 1:10000){
  one_bootstrap_data_frame <- completedat[sample(1:50, 50, replace = TRUE), ]
  bootstrap_dist[i] <- cor(one_bootstrap_data_frame$homicideRate,
  one_bootstrap_data_frame$gunOwnership)
}

corint <- quantile(bootstrap_dist, c(0.025, 0.975))

#Creating correlation interval using the Bootstrap Distribution
bootstrap_dist <- NULL
for (i in 1:10000){
  one_bootstrap_data_frame <- completedat[sample(1:50, 50, replace = TRUE), ]
  bootstrap_dist[i] <- cor(one_bootstrap_data_frame$violentcrimeRate,
  one_bootstrap_data_frame$gunOwnership)
}

corint <- quantile(bootstrap_dist, c(0.025, 0.975))

#Omitted Bootstrap code for difference in two means
obs_stat <- mean(permit_req$violentcrimeRate) - mean(permit_notreq$violentcrimeRate)
combined <- c(permit_req$violentcrimeRate, permit_notreq$violentcrimeRate)
null_dist <- c()
for (i in 1:100000){
  shuff_data <- sample(combined)
  shuff_req <- shuff_data[1:14]
  shuff_notreq <- shuff_data[15:50]
  null_dist[i] <- mean(shuff_req) - mean(shuff_notreq)
}

(p_val <- sum(null_dist <= obs_stat)/100000)

obs_stat <- mean(permit_req$firearmDeathRate) - mean(permit_notreq$firearmDeathRate)
combined <- c(permit_req$firearmDeathRate, permit_notreq$firearmDeathRate)
null_dist <- c()
```

```
for (i in 1:100000){
  shuff_data <- sample(combined)
  shuff_req <- shuff_data[1:14]
  shuff_notreq <- shuff_data[15:50]
  null_dist[i] <- mean(shuff_req) - mean(shuff_notreq)
}

(p_val <- sum(null_dist <= obs_stat)/100000)
```